Translational Oncology
www.transonc.com

Volume 8 Number 1 February 2015 pp. 55–64 55

Determining the Variability of Lesion Size Measurements from CT Patient Data Sets Acquired under "No Change" Conditions[1,2]

Michael F. McNitt-Gray*, Grace Hyun Kim*, Binsheng Zhao†, Lawrence H. Schwartz†, David Clunie‡, Kristin Cohen§, Nicholas Petrick¶, Charles Fenimore#, Z.Q. John Lu and Andrew J. Buckler††**

*David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; †Columbia University Medical Center, New York, NY, USA; ‡Pixel Med Publishing, LLC, Bangor, PA, USA (formerly with Core Lab Partners, Inc, Princeton, NJ); §Janssen Pharmaceutical Research and Development (formerly with Core Lab Partners, Inc, Princeton, NJ), Titusville, NJ, USA; ¶Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, USA; #The Image-Quality Measurement Consultancy, Gaithersburg, MD, USA (formerly with National Institute of Standards and Technology, Gaithersburg, MD); **National Institute of Standards and Technology, Gaithersburg, MD, USA; ††Elucid Bioimaging Inc, Wenham, MA, USA

Abstract

PURPOSE: To determine the variability of lesion size measurements in computed tomography data sets of patients imaged under a "no change" ("coffee break") condition and to determine the impact of two reading paradigms on measurement variability. *METHOD AND MATERIALS:* Using data sets from 32 non-small cell lung cancer patients scanned twice within 15 minutes ("no change"), measurements were performed by five radiologists in two phases: (1) independent reading of each computed tomography dataset (timepoint): (2) a locked, sequential reading of datasets. Readers performed measurements using several sizing methods, including one-dimensional (1D) longest in-slice dimension and 3D semi-automated segmented volume. Change in size was estimated by comparing measurements performed on both timepoints for the same lesion, for each reader and each measurement method. For each reading paradigm, results were pooled across lesions, across readers, and across both readers and lesions, for each measurement method. *RESULTS:* The mean percent difference (± SD) when pooled across both readers and lesions for 1D and 3D measurements extracted from contours was 2.8 ± 22.2% and 23.4 ± 105.0%, respectively, for the independent reads. For the locked, sequential reads, the mean percent differences (± SD) reduced to 2.52 ± 14.2% and 7.4 ± 44.2% for the 1D and 3D measurements, respectively. *CONCLUSION:* Even under a "no change" condition between scans, there is variation in lesion size measurements due to repeat scans and variations in reader, lesion, and measurement method. This variation is reduced when using a locked, sequential reading paradigm compared to an independent reading paradigm.

Translational Oncology (2014) 8, 55–64

Address all correspondence to: Michael McNitt-Gray, PhD, DABR, FAAPM, Professor, Department of Radiological Sciences, 924 Westwood Blvd, Suite 650, David Geffen School of Medicine at UCLA, Los Angeles, CA 90024, USA.
E-mail: mmcnittgray@mednet.ucla.edu

[1]Grants supporting this research: Partial support was provided by a grant from the National Institutes of Health to the Radiological Society of North America (NHLBI-PB-EB-2010-159-JKS "Recovery—Quantitative Imaging Biomarker Alliance").

[2]This article refers to supplementary material, which is designated by Table A1 and is available online at www.transonc.com.

Received 23 September 2014; Revised 7 January 2015; Accepted 13 January 2015

1936-5233/15
http://dx.doi.org/10.1016/j.tranon.2015.01.001

Introduction

Advances in computed tomography (CT) technology over the past several decades have led to exquisitely detailed descriptions of anatomy and pathology and rapidly decreasing doses. The current multidetector row CT systems permit the acquisition of submillimeter thickness image data sets of a patient's entire thorax in a single breath hold. These very highresolution image data sets can also be used with advanced three-dimensional (3D) techniques to visualize and quantify anatomy and pathology with unprecedented detail and accuracy, using relatively noninvasive techniques. This has led to lesion size measurements derived from CT images being heavily used in the assessment of treatment response in the setting of cancer patients.

Despite advances in both CT technology and our understanding of cancer's cellular and molecular mechanisms, the current standard method to measure tumor response to therapy using CT remains the Response Evaluation Criteria in Solid Tumors (RECIST), which is based on unidimensional, linear measurements of tumor diameter [1–3]. Because it is based on a series of linear measurements, RECIST offers a simple approach that requires minimal effort. The RECIST guidelines, however, presume that tumors are spherical and change in a uniform manner. Significant variability in the RECIST measurements exists across different observers [4], and published work generally focuses on the surrogate of "best overall response," with only a few methods addressing other imaging endpoints such as "time to progression" and "disease-free survival." As a therapy response measurement procedure, RECIST maps these linear measurements into one of four categories: complete response, partial response, stable disease, and progressive disease.

Because it uses only unidimensional linear measurements in its assessment, the RECIST criteria do not fully use the much higher resolution data sets offered by modern multidetector row CT systems. This may limit the ability to accurately reflect size changes that occur in the many lesions that are not spherical in nature and may ultimately limit the ability to identify early changes in patients undergoing treatment [4,5]. The advances in CT technology described above had led to the development of 3D methods to estimate the volume of lung lesions, with the aim of developing more accurate and consistent measurements, even for non-spherical lesions, to ultimately better assess response over a shorter time interval. Tumor volumetric measurements may be more sensitive indicators of change and, by inference, response to treatment. However, volume measurements are fundamental estimates that have variability from multiple sources—patient effects, scanner-related effects, measurement and reader effects, and so on. Petrick et al. [6] investigated the use of anthropomorphic phantoms to estimate and compare the bias and variance of measurements of the size of spherical and complex simulated lung nodules using unidimensional, bidimensional, and volumetric measures. They showed that 3D volumetric measurements in phantom lesions demonstrate a smaller bias as well as smaller variability compared to unidimensional measurements, especially for complex lesion shapes.

This effort extends the previous work related to the use of several sizing methods from CT image data as a biomarker of response by reporting on investigations that used "coffee break" CT data sets where patients were scanned twice within a very short (15-minute) period of time using the same imaging technique and on the same scanner [7,8]. These data sets were then analyzed with several size measurements (including unidimensional, bidimensional, and volumetric methods) to assess measurement variability under a "no change" condition. Measurements were performed by five different radiologists under two different reader paradigms described below and allow us to estimate the inherent variability in measuring lesions in patient data sets using several different size measures. These estimates of variability will in turn inform thresholds for determining whether a meaningful change has occurred in tumor size.

The purpose of this work was to quantify the variability of lesion size measurements under a "no change" condition with an independent radiology reading facility performing reads under two distinct reading conditions. The reading was performed under one condition where reads were done independently and one where reads were done with a locked, sequential reading paradigm, which is more reflective of clinical trials practice.

Methods

This study was carried out in two phases, which were defined in terms of the reading paradigm used in each phase. In the first phase ("independent read"), radiologists read each time point of each case independently in random order, and they were not allowed to consult any previous measurements they made on the case. In the second phase ("locked sequential read"), radiologists read the first time point scan, locked their measurements, and then made measurements on the second time point scan while being allowed to review their prior measurements on the first time point scan.

Case Selection and Inclusion Criteria

The first phase of this study used the "no change" cases from the publicly available Reference Image Database to Evaluate Response to Therapy (RIDER) database [9] that were originally performed at the Memorial Sloan Kettering Cancer Center [7,8]. This data set consisted of 32 non-small cell lung cancer patients who were scanned twice on the same scanner within 15 minutes. Each data set consisted of a thoracic CT scan performed without the use of any iodinated contrast agent and reconstructed with thin section (1.25 mm) images.

For these data sets, only one lesion per patient was selected for measurement (32 lesions total), with each lesion visible on each of the two time point (time point 1 and time point 2) scans. The approximate lesion diameters ranged from 8 to 40 mm. The shapes of the selected lesions ranged from simple and isolated to complex and cavitated. A few sample lesions are illustrated in Figure 1.

For the second phase of this study, a set of distractor lesions was identified from the RIDER data set. The purpose of these distractor lesions was to reduce bias in the reader study by having radiologists also measure lesions that change. These distractor cases would reduce the potential for readers to recognize or assume that every case they were measuring was a "no change" case. Therefore, these cases were not from the "coffee break" data set but were selected from other RIDER databases containing thoracic CT scans of non-small cell lung cancer patients performed without the use of any iodinated contrast agent. Twenty lesions were selected from patients who had multiple scans (even if they were months apart), who were reconstructed with thin sections (≤1.25 mm thickness), and who demonstrate some amount of change in the CT-based lesion size across the two selected scans.

In this study, only the coffee break lesions were analyzed; the distractor change cases served only to control potential reader bias. The analysis cases were retrospectively divided into two types as follows: 1) lesions that were classified as meeting the conditions described in the "Claims" section of the "Quantitative Imaging Biomarker Alliance (QIBA) Profile: CT Tumor Volume Change (CTV-1)" [10] and 2) lesions that did not meet

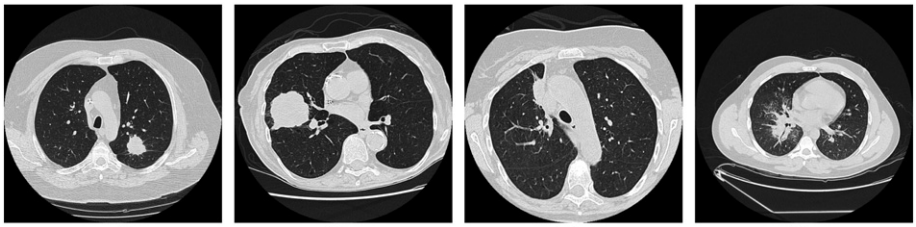

Figure 1. Examples of lesions used: (A) and (B) are examples of lesions that were judged to have met the QIBA CTV-1 profile claim [10], while (C) and (D) are examples of lesions that were judged to have not met the QIBA CTV-1 profile claim.

these conditions. Specifically, the claims section of the QIBA CTV-1 profile states that the claims are only applicable "when the given tumor is measurable (i.e. tumor margins are sufficiently conspicuous and geometrically simple enough to be recognized on all images….) and the longest in-plane diameter of the tumor is 10 mm or greater." Therefore, lesions described as meeting the QIBA CTV-1 profile were those that were judged to have clearly identified tumor margins by one of the authors (M.F.M.-G.) who was also part of the QIBA profile development team; all coffee break lesions used in this study exceeded the 10-mm diameter threshold.

Readers

For phase 1, five readers performed the measurements described below. The number of readers was based on recommendations [11,12] and available resources. For phase 2, there were also five readers, three of whom participated in phase 1. All readers were board-certified radiologists with experience as readers for multicenter oncology clinical trials. All were experienced with performing linear measurements as part of the RECIST assessment protocol.

Reading Software

The cases were reviewed and the lesion size measurements were made using a modified software package that the readers were familiar with from oncology readings performed as part of their normal workload. Therefore, all readers could use the software to load, view, and measure lesions for single longest diameter measurements and bidirectional diameter measurements. For volumetric measurements, readers were trained on the modified software that allowed them to create and edit contours of each lesion. From the contours, lesion volume was calculated. In addition, the single longest diameter and bidirectional diameters were derived from the lesion contours.

Measurements

In phase 1 of the study, each radiologist measured each lesion in three different ways as described below. For this phase, each measurement was made independently in a separate reading session, so the reader did not have access to any of their prior measurements for reference:

- 1D: manual linear measurements using a Caliper tool to obtain the single longest diameter on one image.
- 2D: manual bidirectional diameters using the Orthogonal Ruler tool to obtain the longest diameter and diameter perpendicular to the longest diameter. Because this measurement was done independently, the single longest diameter was also retained for comparison.
- 3D: algorithm assisted volume using the boundary contour tool. These volumetric measurements were made using software based on a prototype proprietary semiautomated tool (Oncocare Prototype; Siemens Corporate Research, Princeton, NJ), which included a lesion segmentation component [13,14]. The 3D measurement process was given as follows: the reader 1) defined a seed stroke across the lesion (i.e., a RECIST-like line across the perceived maximum diameter of the lesion), 2) applied the segmentation tools, 3) evaluated the quality of the segmentation, and 4) refined or added seed strokes and reapplied the segmentation tool until satisfied with the 3D nodule segmentation. From the resulting contour, the volume was calculated as well as both the single longest diameter in a given image and the diameter perpendicular to that longest diameter.

In phase 2, using the same boundary contour tool described above, the readers were asked to obtain the entire lesion boundary contour for both the no change lesions and the distractor lesions. From each contour, the 3D volume values were obtained; in addition, the 1D single longest diameter and 2D longest and perpendicular diameter values were automatically derived from the same contour information.

Reading Experiments

As mentioned above, the reading experiments were carried out in two distinct phases.

Phase 1— Independent readings. In this phase, readers were shown only the original 32 "no change" lesions described above. The lesion identified in each time point scan was assessed in a different session. Readers were *not* allowed to see any of their previous measurements (to ensure that *each measurement was done independently*). Measurements were performed using the three methods described above, but each measurement type was performed in a different session. The order of reading was randomized by patient, scan (time point 1 and time point 2), and measurement type. Reading order was different for each reader as well.

Phase 2— Locked sequential readings. In comparison to the independent reading paradigm of phase 1, this phase allowed the readers to perform a "locked, sequential read" that allow the reader to see the measurements they made on the previous time point scan. In this phase, due to time (and budget) constraints, the readers only provided the contours of the entire lesion boundary. From these contours, lesion volume and longest diameter were estimated. To prevent possible bias in this reading paradigm associated with the readers recognizing or assuming that all of the cases were "no change" data sets in this phase, the readers were not told the aims of the study and were shown both the original 32 "no change" lesions described above as well as the 20 distractor lesions. It should be noted that *only* the measurements performed on the "no change" lesions were used in analysis and it is assumed their volume change across the two time points was zero.

The locked, sequential read was accomplished by 1) first showing the reader one time point scan (randomly selected time point 1 or time point 2), having the reader contour the entire lesion boundary, locking that result (i.e., no editing allowed), and then 2) immediately showing the reader the other time point scan, allowing the reader to see the contour on the previously measured lesion and then having the reader contour the entire lesion on the second scan. No copy/paste of lesion contours was allowed. In this phase, the order of reading was randomized only by patient. As in phase 1, the reading order was different for each reader.

Analysis

Size estimates from six different size measurements were provided by each of the five readers in the phase 1 independent reads: 1D longest diameter from the caliper, from the orthogonal ruler, and from the contoured lesion, 2D area measurement defined as the product of the longest and perpendicular diameters from the orthogonal ruler and the contoured lesion, and a volume estimate from the contoured lesion. For the second reading phase of locked, sequential reads, three size measurements were provided by the second set of five readers: the 1D, 2D, and volume from a contoured lesion. The percent change between the two time point scans was calculated as the ratio of the difference in the second read and the first read to the first read. The proportional change was calculated as the ratio of the difference between the second read and the first read to the average between the first and second reads. Descriptive summary statistics and changes were reported for each phase.

$$\text{Percent change}(\%) = \frac{\text{a metric in the second scan} - \text{a metric in the first scan}}{\text{a metric in the first scan}} \times 100 \quad (1)$$

$$\text{Proportional change}(\%) = \frac{\text{a metric in the second scan} - \text{a metric in the first scan}}{\text{mean of metrics in the two scans}} \times 100 \quad (2)$$

The 95% confidence interval (CI) of the mean and SD were derived from estimation of 1000 bootstrap replications of the observed data and were reported with a Bonferroni adjustment for multiple comparisons. Bland Altman plots [15] were used to show the repeatability in measurements between two scans by an individual reader. A mixed effect model [16] was used to test the differences in percent changes between independent reading and locked sequential read, where a fixed covariate is a dichotomized QIBA type (a QIBA CTV-1 compliant lesion or a non–QIBA CTV-1 compliant lesion as described in [10]) and random effects are subjects and readers nested into a subject.

Box-and-whisker plots are used to show the distribution in the percent change in volume by a QIBA-compliant group *versus* a non-compliant group over the two reading phases. A line within a box indicates the median. The lower hinge and upper hinge indicates 25th and 75th percentiles of data (i.e., interquartile range, IQR). The lower and upper adjacent lines connected to the boxes are 1.5 times of the range of 75th and 25th percentiles. Likewise, the dots are outside values either smaller or greater than the 25th percentile − 1.5 times the IQR or 75th percentile + 1.5 times the IQR, respectively.

A subgroup analysis was also performed with two subgroups: 1) the 20 lesions that were identified as meeting the QIBA CTV-1 profile language (as described in the Case Selection and Inclusion Criteria section above) and 2) the 12 lesions that did not meet the QIBA CTV-1 profile language. The results of each subgroup were provided as well as the pooled results.

In all analyses, a threshold of .05 for a P value was considered to be significant and a threshold between .05 and .1 was considered to show a trend.

Results

Table 1 provides the mean size measurement results, the mean change in size, and the mean percent changes for the independent reading (phase 1) results. These results are stratified by measurement type and pooled across readers and lesions. It should be noted that for consistency in units and to avoid reporting very large numbers, the volume results in Table 1 are given in units of 1000 mm^3. These results indicate that for the independent reading paradigm, the mean percent change for 1D measurements is quite low (mean value of <6%), while the mean percent change for 3D volumetric measurements can be substantially higher (mean value approximately 25%). This table also shows that the mean and SDs of the 1D results are not very different between the various methods of obtaining the longest diameter (direct manual measurement, obtained from perpendicular diameters or extracted from the volumetric contour). Finally, the table shows that both percent changes and the SDs were a bit larger for the 2D methods than those for the 1D methods.

Table 2 provides similar mean size results plus the mean change in size and mean percent changes for the locked sequential reading (phase 2) results. These results are similarly stratified by measurement type and pooled across readers and lesions. These results indicate that for the locked sequential reading paradigm, the mean percent change for 1D measurements is even lower than those in Table 1, showing a mean percent change of only 2.5%. In this reading paradigm, the mean percent change for 3D volumetric measurements were also substantially reduced, demonstrating a mean value of only 7.4% and 2D measurements demonstrate a mean percent change of only 2% (P = .063 and P = .111, retrospectively). The SDs of percent changes are also much smaller here than in the independent reading paradigm reported in Table 1.

Subgroup Analyses

Table 3 shows the mean percent change as well as that of the proportional percent change and the 95% CIs for the independent phase 1 readings broken down by 1) lesions that were identified as being compliant with the QIBA CTV-1 profile, 2) lesions that were identified as not being compliant with the QIBA CTV-1 profile, and also 3) the pooled results across all lesions. Table 4 shows the SDs of percent change and proportional percent (and 95% CIs) for the independent readings (phase 1), also broken down by lesion category and sizing method. These results indicate that the mean percent

Table 1. Mean Value Results (and SD), Differences, and Percent Changes for the Independent Reading (Phase 1) for Each of the Six Sizing Methods Used. Results Are Pooled across Readers and Lesions

Sizing Method	Independent Reading (N = 160 = 32 × 5 Readers)			Percent Changes (%)	Proportional Changes (%)
	First Scan	Second Scan	Difference		
	Mean ± SD (mm, mm^2, mm^3)	Mean ± SD (mm, mm^2, mm^3)	Mean ± SD (mm, mm^2, mm^3)	Mean ± SD	Mean ± SD
1D caliper, mm	33.1 (±19.8)	33.9 (±19.6)	0.8 (±5.1)	5.8 (±23.8)	3.6 (±19.2)
1D orthogonal ruler, mm	32.8 (±20.0)	33.8 (±19.8)	1.0 (±4.3)	5.4 (±21.2)	3.6 (±16.6)
1D from contour, mm	34.2 (±20.1)	34.0 (±19.1)	−0.2 (±5.2)	2.8 (±22.2)	0.9 (±17.6)
2D orthogonal ruler, mm^2	1009 (±1204)	1060 (±1217)	51 (±211)	15.2 (±68.5)	6.2 (±29.6)
2D from contour, mm^2	1098 (±1393)	1067 (±1272)	−31 (±357)	12.7 (±69.0)	2.0 (±35.4)
3D, 1000mm^3	22.14 (±36.2)	22.11 (±35.0)	−0.03 (±7.7)	23.4 (±105)	5.5 (±39.2)

Table 2. Mean Value Results (and SD), Differences, and Percent Changes for the Locked Sequential Reading (Phase 2) for Each of the Three Sizing Methods Used. Results Are Pooled across Readers and Lesions

Sizing Method	Locked Sequential Read (N = 160 = 32 × 5 Readers)			Percent Changes (%)	Proportional Changes (%)
	First Scan	Second Scan	Difference		
	Mean ± SD (mm, mm^2, mm^3)	Mean ± SD (mm, mm^2, mm^3)	Mean ± SD (mm, mm^2, mm^3)	Mean ± SD	Mean ± SD
1D from contour, mm	34.4 (±20.5)	34.8 (±20.4)	0.4 (±3.6)	2.5 (±14.2)	1.7 (±12.0)
2D from contour, mm^2	1156.5 (±1439)	1155.3 (±1454)	−1.2 (±307)	2.1 (±21.2)	0.1 (±18.9)
3D, 1000mm^3	22.9 (±37.4)	22.8 (±37.3)	−0.06 (±5.4)	7.4 (±44.2)	2.2 (±25.5)

Table 3. Estimates of Mean Percent Change and Proportional Change (95% CIs in Brackets) as a Function of Nodule Category and Sizing Method for the Independent Reading (Phase 1).

Phase 1—Independent Read Mean Percent Change							
Nodule Category	Sizing Method	Mean Percent Change (%) [†] and 95% CI		Standard Error	Mean Proportional Change (%) [†] and 95% CI		Standard Error
QIBA profile compliant (N = 100; 20 lesions and five readers)	1D caliper	6.95	[−0.72, 14.61]	2.55	4.28	[−2.11, 10.69]	2.13
	1D ruler	7.16	[0.60, 13.71]	2.18	5.26	[0.40, 10.12]	1.62
	1D contour	4.92	[−2.24, 12.08]	2.38	2.82	[−2.38, 8.02]	1.73
	2D ruler	19.40	[−3.73, 42.54]	7.69	8.55	[0.02, 17.08]	2.83
	2D contour	13.48	[−8.90, 35.86]	7.44	3.12	[−6.20, 12.45]	3.10
	3D volume	30.31	[−10.13,70.76]	13.45	9.05	[−0.80, 18.86]	3.26
QIBA profile non-compliant (N = 60; 12 lesions and five readers)	1D caliper	3.98	[−3.78, 11.76]	2.54	2.31	[−4.16, 8.76]	2.11
	1D ruler	2.54	[−4.80, 9.88]	2.40	0.94	[−5.83, 7.71]	2.21
	1D contour	−0.88	[−7.94, 6.18]	2.31	−2.41	[−9.38, 4.55]	2.28
	2D ruler	8.23	[−8.96, 25.44]	5.62	2.38	[−9.54, 14.30]	3.90
	2D contour	11.43	[−12.56, 35.44]	7.84	0.15	[−16.93, 17.23]	5.58
	3D volume	16.12	[−16.39, 48.63]	10.62	−0.27	[−19.05, 18.52]	6.14
All (N = 160; 32 lesions and five readers)	1D caliper	5.84	[0.31, 11.36]	1.86	3.55	[−0.95, 8.04]	1.51
	1D ruler	5.42	[0.37, 10.47]	1.69	3.64	[−0.24, 7.52]	1.30
	1D contour	2.75	[−2.34, 7.83]	1.71	0.86	[−3.22, 4.93]	1.37
	2D ruler	15.21	[−0.85, 31.28]	5.39	6.24	[−0.52, 12.99]	2.27
	2D contour	12.71	[−3.41, 28.84]	5.41	2.01	[−6.43, 10.44]	2.83
	3D volume	23.40	[−2.36, 52.34]	9.18	5.55	[−3.68, 14.78]	3.10

[†] 95% CIs, based on t-distribution with bootstrap within each subgroup and adjusted using a Bonferroni correction for comparisons of all six types of sizing methods, are shown in brackets.

change and the SD of the percent change are similar between lesion types (QIBA CTV-1 compliant or non–QIBA CTV-1 compliant) in the independent reading paradigm; they are not statistically significantly different (P = .469 and P = .719, respectively). This appears to be true regardless of the sizing method (1D, 2D, 3D) used. There are differences between the methods themselves, but these differences do not change depending on the nodule category. In addition, these results show that for the independent reading

Table 4. Estimates of SD of the Percent Change and Proportional Change (95% CIs in Brackets) as a Function of Nodule Category and Sizing Method for the Independent Reading (Phase 1)

Phase 1—Independent Read SD Percent Change							
Nodule Category	Sizing Method	SD of Percent Change (%) [†] and 95% CI		Standard Error	SD of Proportional Change (%) [†] and 95% CI		Standard Error
QIBA profile compliant (N = 100; 20 lesions and five readers)	1D caliper	25.79	[11.34, 40.24]	4.92	20.56	[11.91, 29.20]	2.95
	1D ruler	22.02	[6.99, 37.04]	5.12	16.18	[8.35, 24.02]	2.67
	1D contour	24.17	[9.13, 39.21]	5.12	17.46	[9.10, 25.81]	2.85
	2D ruler	79.65	[1.68, 157.62]	26.56	29.22	[15.19, 43.25]	4.78
	2D contour	73.94	[7.76, 140.11]	22.55	31.45	[17.26, 45.65]	4.84
	3D volume	102.21	[30.28, 174.13]	39.09	33.11	[15.57, 50.67]	5.98
QIBA profile non-compliant (N = 60; 12 lesions and five readers)	1D caliper	20.21	[8.02, 32.39]	4.15	16.70	[9.07, 24.33]	2.60
	1D ruler	19.52	[7.13, 31.91]	4.22	17.00	[8.18, 25.83]	3.01
	1D contour	18.01	[9.21, 26.81]	3.00	17.56	[9.45, 25.67]	2.76
	2D ruler	43.63	[3.82, 83.44]	13.56	29.93	[15.90, 43.96]	4.80
	2D contour	60.52	[25.44, 95.61]	11.95	41.39	[25.94, 56.84]	5.26
	3D volume	110.02	[54.08, 165.96]	19.21	47.76	[28.57, 66.95]	6.54
All (N = 160; 32 lesions and five readers)	1D caliper	23.83	[13.47, 34.19]	3.53	19.17	[12.89, 25.44]	2.14
	1D ruler	21.17	[10.37, 31.97]	3.68	16.58	[11.14, 22.01]	1.85
	1D contour	22.18	[11.48, 32.88]	3.65	17.62	[11.65, 23.60]	2.04
	2D ruler	68.45	[11.58, 125.32]	19.37	29.55	[19.12, 39.97]	3.55
	2D contour	69.02	[24.73, 113.31]	15.09	35.41	[24.96, 45.86]	3.56
	3D volume	105.01	[62.89, 147.30]	14.38	39.37	[28.89, 49.85]	3.57

[†] 95% CIs, based on bootstrap t statistic using 1000 replications within each subgroup and adjusted using a Bonferroni correction for comparisons of all six types of sizing methods, are shown in brackets. Due to the random sampling of the bootstrapping, it is not necessarily expected that the 95% CIs will be symmetric about the SD.

Table 5. Estimates of Mean Percent Change and Proportional Change (95% CIs in Brackets) as a Function of Nodule Category and Sizing Method for the Locked, Sequential Reading (Phase 2)

Phase 2—Locked, Sequential Read Mean Percent Change							
Nodule Category	Sizing Method	Mean Percent Change (%)[†] and 95% CI		Standard Error	Mean Proportional Change (%) and 95%CI		Standard Error
QIBA profile compliant (N = 100; 20 lesions and five readers)	1D contour	2.42	[−2.41, 7.24]	1.60	1.40	[−2.45, 5.26]	1.28
	2D contour	0.66	[−6.01, 7.32]	2.22	−1.35	[−7.39, 4.70]	2.01
	3D volume	8.32	[−6.60, 23.23]	4.96	2.09	[−5.49, 9.66]	2.52
QIBA profile non-compliant (N = 60; 12 lesions and five readers)	1D contour	2.69	[−1.77, 7.15]	1.46	2.09	[−1.97, 6.15]	1.33
	2D contour	4.40	[−3.53, 12.33]	2.59	2.56	[−4.43, 9.55]	2.29
	3D volume	5.92	[−6.07, 17.93]	3.92	2.36	[−7.27, 2.00]	3.15
All (N = 160; 32 lesions and five readers)	1D contour	2.52	[−0.28, 5.33]	1.16	1.66	[−0.68, 4.01]	0.97
	2D contour	2.06	[−2.12, 6.24]	1.73	0.12	[−3.41, 3.65]	1.46
	3D volume	7.42	[−0.98, 15.82]	3.47	2.19	[−2.52, 6.90]	1.95

[†] 95% CI, based on t-distribution within each subgroup and adjusted using a Bonferroni correction for comparisons of the three types of sizing methods, are shown in brackets.

Table 6. Estimates of SD of the Percent Change and Proportional Change (95% CIs in Brackets) as a Function of Nodule Category and Sizing Method for the Locked, Sequential Reading (Phase 2)

Phase 2—Locked, Sequential Read SD Percent Change							
Nodule Category	Sizing Method	SD of Percent Change (%)[†] and 95% CI		Standard Error	SD of Proportional Change (%) and 95%CI		Standard Error
QIBA profile compliant (N = 100; 20 lesions and five readers	1D contour	15.82	[7.17, 24.46]	2.94	12.89	[6.99, 18.79]	2.01
	2D contour	21.90	[10.26, 33.53]	3.96	19.14	[12.56, 25.73]	2.24
	3D volume	50.64	[9.07, 92.21]	14.16	26.47	[14.15, 38.79]	4.20
QIBA profile non-compliant (N = 60; 12 lesions and five readers)	1D contour	11.07	[7.00, 15.14]	1.39	10.48	[7.02, 13.94]	1.18
	2D contour	19.82	[12.77, 26.87]	2.40	18.45	[13.21, 23.69]	1.79
	3D volume	30.95	[10.36, 51.54]	7.02	23.97	[13.25, 34.70]	3.65
All (N = 160; 32 lesions and five readers)	1D contour	14.19	[8.02, 20.35]	2.10	12.01	[8.06, 15.97]	1.35
	2D contour	21.16	[13.47, 28.85]	2.62	18.92	[14.52, 23.32]	1.50
	3D volume	44.20	[15.43, 72.97]	9.80	25.48	[16.78, 34.19]	2.97

[†] 95% CI, based on t-distribution within each subgroup and adjusted using a Bonferroni correction for comparisons of the three types of sizing methods, are shown in brackets.

paradigm, the mean and SD percent change values are higher for volume (3D) than for perpendicular diameters (2D), which are higher than single diameters (1D).

Table 5 shows the mean values of the percent change, the proportional percent change, as well as 95% CIs for the locked sequential readings (phase 2) broken down by 1) lesions that were identified as being compliant with the QIBA CTV-1 profile, 2) lesions that were identified as not being compliant with the QIBA CTV-1 profile, and also 3) the pooled results of all lesions. Table 6 shows SDs of percent change and proportional percent (and 95% CIs) for the locked sequential readings (phase 2), also broken down by lesion category and sizing method. These results indicate that the mean percent change and the SD of the percent change are similar between lesion types (QIBA CTV-1 compliant or non–QIBA CTV-1 compliant) in the locked sequential reading paradigm; they are not statistically significantly different (P = .739 and P = .940, respectively). This appears to be true regardless of the sizing method (1D, 2D, 3D) used. In addition, Table 5 shows that for this locked sequential reading paradigm, the mean percent change values are now much more similar between sizing methods and that the proportional change values are nearly identical across the lesion types for all of the sizing methods. Table 6 does show that there are still differences in the SDs between sizing methods, with volume (3D) having a larger SD than for perpendicular diameters (2D), which are in turn higher than single diameters (1D).

Figure 2 illustrates the effect of the reading paradigm (independent reading of phase 1 *vs* the locked sequential reading in phase 2) on the percent change for the two categories of lesions identified in this study (QIBA CTV-1 compliant *vs* non–QIBA CTV-1 compliant). This figure shows that in phase 1, the median (interquartile range) in percentage changes is 0.7% (18.6%) for QIBA CTV-1 compliant lesions and 2.8% (32.9%) for the non-compliant type. In phase 2, median (interquartile range) in percentage changes are −0.8% (20.9%) for the QIBA CTV-1 compliant type and 2.7% (16.3%) for the non-compliant type.

For lesions that would not be considered to be compliant with the QIBA CTV-1 profile, the interquartile range has been reduced in half

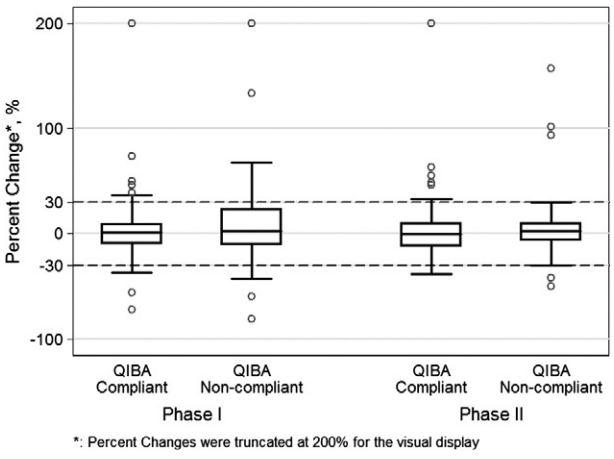

Figure 2. Box plots of volume percent change by lesion category and reading phase (phase 1—independent reads; phase 2—locked sequential reads). The dashed lines represent the ±30% change values described in the QIBA CTV-1 profile.

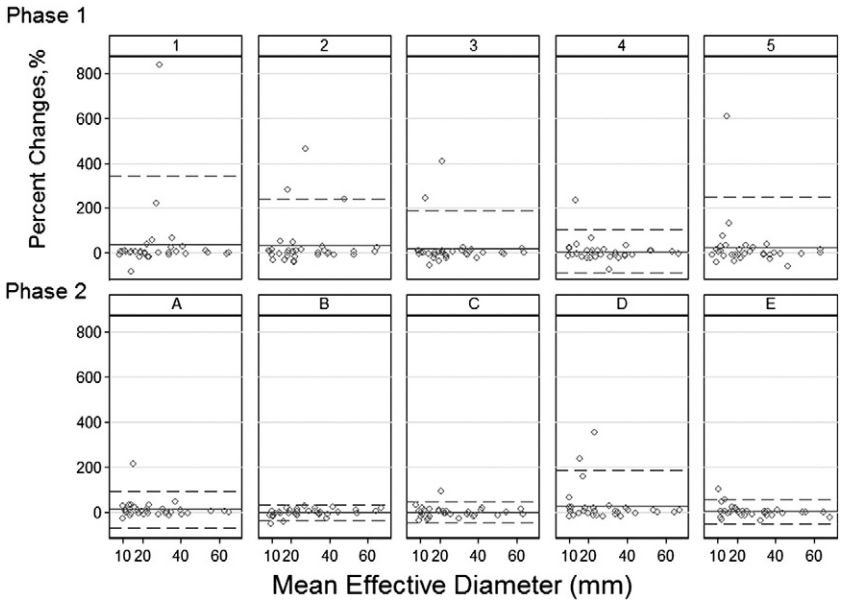

Figure 3. Bland-Altman plots of volume percent change by reading phase (phase 1—independent reads; phase 2—locked sequential reads) for individual readers, plotted as a function of lesion volume (in 1000 mm^3). The dashed lines represent 1 SD from the mean. It should be noted that the readers in phase 1 were not the same five readers as in phase 2 (hence, readers are identified by number in phase 1 and by letter in phase 2).

by following the locked sequential reading paradigm (phase 2) when compared with an independent reading paradigm (phase 1). Within these QIBA non-compliant lesions, a trend toward reducing the percent error was found between phases by the mixed effect model, but this difference was not statistically significant ($P = .069$). Within the QIBA CTV-1 compliant data set, there was no statistically significant difference in going from independent reading to the locked sequential reading based on percent error ($P = .363$).

This figure also demonstrates that for the QIBA CTV-1 compliant lesions, the percentage of observations in which the volume percent change falls within ± 30% (as stated in the QIBA CTV-1 profile claim language) is 84.0% (84 of 100 observations – 20 lesions × 5 reviewers) for phase 1, which increases to 86.0% for the phase II reading paradigm. For the non–QIBA CTV-1 compliant lesions, the percentage of observations in which the volume percent change falls within ± 30% is 71.7% (43 of 60 observations – 12 lesions × 5 reviewers) for phase 1, which increases to 91.7% for the phase 2 reading paradigm. Over the entire population of both compliant and non-compliant lesions, the percentage of observations where the volume percent change falls within ± 30% is 79.4% (127 of 160 observations – 32 lesions × 5 reviewers) for phase 1, which increases to 88.1% for the phase 2 reading paradigm.

Figure 3 shows the Bland-Altman plots of volume percent change for each reading paradigm for each individual reader. This figure illustrates the differences between individual readers in each phase and emphasizes that smaller lesions were most likely to have larger percent differences. This figure also illustrates the differences between readers, especially in the phase 1 independent reading paradigm, but there are still instances of large differences in the phase 2 locked sequential readings. Specifically, in phase 1, the mean (± 2 * SD) percentage changes are 36.5% (– 269.3%, 342.3%), 32.1% (– 174.7%, 238.9%), 17.9% (– 151.7%, 187.5%), 6.2% (– 89.6%, 102%), and 24.4% (– 199.6%, 248.4%) for reader 1, reader 2, reader 3, reader 4, and reader 5, respectively. In phase 2, mean (± 2 * SD from Bland-Altman plot (15)) percentage changes are 11.8% (– 69.4%, 93%), – 2.8% (– 36.4%, 30.8%), – 1.4% (– 49%, 46.2%), 27% (– 130.6%, 184.6%), and 2.6% (– 49.8%, 55%) for reader A, reader B, reader C, reader D, and reader E, respectively.

Comparing the two phases, there is a trend toward a reduction in the mean percent change in volume ($P = .063$). However, a similar trend was not found for 1D and 2D ($P = .99$ and $P = .11$, respectively). Nodule category (QIBA CVT-1 compliant or not) was not statistically significant for any of the three sizing methods ($P = .22$, $P = .14$, and $P = .61$, respectively).

Discussion

The purpose of this study was to determine the variability in lesion size measurements for patients imaged on CT under a "no change" condition. This is an important step in investigating the use of changes in CT image–based estimates of tumor size as a biomarker of response. If estimates of tumor size are to serve as more sensitive indicators of change (and hence response to treatment) than current methods, then the sources and magnitude of variation need to be understood under conditions encountered in clinical practice. This effort extends previous work by investigating the variation that exists under a "no change" condition using two different reading paradigms, three different sizing methods, five different readers, and 32 lesions imaged by two scans performed within 15 minutes.

Our results indicate that the reading paradigm did make a substantial (but not statistically significant) difference in the measurement variability. The locked sequential reading paradigm provided much lower mean differences and much lower SDs than the independent reading paradigm performed in phase 1 as shown in Figure 2. In fact, Figure 2 shows that there is a trend toward improvement and the largest improvement is in non–QIBA CTV-1 compliant lesions, where the interquartile range decreased by half. Though the improvement in QIBA compliant lesions is not large, the locked sequential read performs no worse than the independent reading paradigm.

In the locked sequential reading paradigm, radiologists annotate the lesion imaged at one time point, and when they are finished, then that annotation is "locked" (no further editing allowed) and then the second time point is displayed and annotated. The radiologist is allowed to consult the annotation performed on the first time point. As one might expect, this leads to much more consistent annotation of lesion boundaries. Therefore, one of our strongest recommendations is to perform the locked sequential reading paradigm wherever possible.

Readers were shown to be a source of variability based on the limit of agreements from Bland-Altman plot within each phase. Figure 3 demonstrates that even within a reading paradigm, there are noticeable differences in measurement variability between readers. This is especially true for the independent reading paradigm. Perhaps the largest improvement is due to the locked sequential reading paradigm that allowed readers to be more consistent with their own annotations by allowing them to refer to their previous markings as they annotate the subsequent time point scan. This is most likely the cause of the reduced variability when using the locked sequential reading paradigm.

The change metric can also make a substantial difference. Two metrics were described—percent change and proportional change [in equations (1) and (2) in the Analysis section]. The proportional change has been suggested as a metric that is less biased by lesion size and has been recommended for use by the QIBA Metrology Working Group [17,18]. The results from the current work demonstrate that the proportional change consistently provides a lower mean value of change as well as lower SDs. One example to illustrate these sources of variability is described in Appendix A.

This work shows that variability within a sizing method may be influenced by the reading paradigm. For example, the 1D sizing method results do not change significantly or substantially across reading paradigms (of note, the means were 2.75% with 95% CI [−2.34%, 7.83%] in the independent reading and 2.52% with 95% CI [−0.28%, 5.33%] in the locked sequential reading). However, volume measurements do change substantially and differences are lower for the locked sequential reading paradigm, but this did not reach statistical significance (P = .067; of note, the summary statistic, the means were 23.40% with 95% CI [−2.36%, 52.34%] in the independent reading and 7.42% with 95% CI [−0.98%, 15.82%] in the locked sequential reading). This may be due to the number of choices each radiologist has to make when contouring the boundary of an entire lesion (i.e., the reader has to define every voxel to include as part of the lesion) and the consistency of those choices when being able to refer to a previously generated contour. When making a single linear measurement, there are fewer choices to make on an individual lesion (i.e., the reader has to only define the slice and the endpoints of the longest dimension) and so variability may not be as affected by the different reading paradigms.

The results of this work do show higher percent change values, even for the locked sequential reading paradigm, for the 3D (volumetric) measurements than for the 1D (linear) measurements across all lesions: 7.4% for mean percent change in 3D compared to 2.5% mean percent change in 1D; 2.2% for mean proportional change in 3D compared to 1.7% mean proportional change for 1D. However, Petrick [6] suggested normalizing percent size change metrics to a common scale to allow comparisons. This is accomplished by converting 2D and 3D size estimates to a 1D (mm) scale by taking the square root and cube root, respectively, of the size estimates. While in this manuscript we reported the raw (unnormalized by size) percent and proportional differences in the tables, the normalized 3D (volumetric) values would be 2.4% for mean percent change and 0.7% for the mean proportional change, which are quite comparable to the 1D (linear) values. Therefore, our conclusion is similar to that of Petrick, which is that 3D and 1D can both provide small measurement variability and especially so when the locked sequential read is used. Unlike Petrick's work, this work is unable to assess bias as the true lesion size is not known.

Finally, this work did not show much difference in measurement variability across different lesion categories (QIBA CTV-1 profile compliant or non-compliant; P = .363). That profile limited its claims to those lesions where "tumor margins are sufficiently conspicuous and geometrically simple enough to be recognized on all images." The results in Tables 3 to 6 demonstrated no significant differences in mean percent change, SD of percent change, mean proportional change, or SD of proportional change across lesion category for any of the sizing methods or for either reading paradigm. Results from Figure 2 indicate that there are both QIBA compliant and non-compliant lesions that can result in outliers that produce significant variation in size (>200% difference in volume). However, this figure indicates that for the independent reading paradigm performed in phase 1, the IQR for the QIBA non-compliant lesions exceeds the ±30% claim of the QIBA CTV-1 profile. It also shows that for the locked, sequential reading paradigm there is small change in the mean percent differences but some improvement in the outliers for both QIBA compliant and non-compliant lesions as well as theIQR for the non-compliant lesions. Although the 95% CIs of Tables 3 to 6 indicate that these differences are not statistically significant, this is perhaps due to the small sample size of lesions (20 in the QIBA CTV-1 compliant group and only 12 in the non-compliant group). The overall mean differences between the two phases were 16% (=23% − 7%) with 32 subjects and 5 readers (Tables 3 and 5). These results do indicate a trend toward improvement, both in reducing variance and in reducing the number and range of the outliers, when the locked sequential read is used, especially for non-compliant lesions. The result that only a trend (i.e., differences did not reach statistical significance) was found between the two reading paradigms can be due to several factors, including performance of the experienced radiologists, whose markings of volume segmentations reached large variations in only a few cases in the independent reading. A simulation study was performed, which indicated that a sample size of 41 subjects was required to achieve approximately 80% power to detect a difference of 16% in volume between the two phases at the 5% significance level. This was based on having a paired reading with a correlation of 0.5 and using a two-sided paired t test based on 20000 Monte Carlo samples using the mean and SD values from Tables 3 and 5 [19].

A limitation of our study is that we do not know clinical truth for the individual tumors, only that the difference between scans should be zero. Therefore, we cannot determine the accuracy for any of the individual measurements. This limitation is not unique to our study but is a limitation of any study that analyzes patient CT data. However, analysis of the accuracy and precision, or their combination, is likely important to fully characterize the performance of a quantitative imaging biomarker. It is also clear from our study that the differences observed between scans of the same lesion are unlikely to be solely associated with inconsistencies produced by the software tool and reader. Instead, some lesions actually present differently in the two CT scan data such that simply segmenting the lesions would

Table A1. Results for the Case Shown in Figure 1*D* Using Volume Measurements and Percent Change Calculations for Each Reader, Reported for Both the Independent Reading and Locked Sequential Reading Phases.

Reading Paradigm	Reader	Volume 1 (mm^3)	Volume 2 (mm^3)	Percent Change (%)	Proportional Change (%)
Phase 1—independent read	1	5,220	16,720	219	105
Phase 1—independent read	2	25,750	87,700	240	109
Phase 1—independent read	3	15,720	17,060	8.49	8.15
Phase 1—independent read	4	17,910	14,980	−16.4	−17.8
Phase 1—independent read	5	10,340	12,890	24.7	22.0
Phase 2—locked sequential read	A	17,630	16,690	−5.31	−5.45
Phase 2—locked sequential read	B	17,370	18,210	4.84	4.72
Phase 2—locked sequential read	C	16,090	11,730	−27.1	−31.3
Phase 2—locked sequential read	D	13,140	16,760	27.5	24.2
Phase 2—locked sequential read	E	25,520	21,760	−14.7	−15.9

systematically show a false change in lesion size. This type of variability likely defines a lower bound on achievable quantitative performance for a particular acquisition protocol.

Although this study used the same CT image data that were used in the study described by Zhao et al. [7], there were several important differences between these two studies including the selection of different lesions, different readers and numbers of readers (three in Zhao et al. and five in this study), and a longer washout period between readings for this study. It should also be noted that in this study two different reading phases (independent and locked, sequential reading) were used, while in the study of Zhao et al. a sequential reading paradigm was used. These differences make direct comparisons difficult; however, both studies have shown that variation increases for smaller sized lesions, independent of the sizing method used.

Quantitative imaging–based biomarkers involve more than just the acquisition and interpretation of image data; they require consistent application of protocols to reduce both bias and variance of a measurement system. For a tumor size–based biomarker using CT imaging, this means that these protocols must address many aspects including patient preparation, image acquisition and reconstruction, image analysis, interpretation paradigm, and even inclusion criteria (size, complexity, and so on) for lesions to be measured. Each of these factors contributes to the bias and variance of the measurement system; therefore to use imaging biomarkers as a sensitive measure of treatment response requires addressing each of these issues. In this work, we have addressed some of the important and necessary issues in reading paradigm, lesion size, and complexity, as well as measurement method in making CT tumor size a viable imaging biomarker.

Acknowledgements

The authors acknowledge the efforts of several key contributors to this work. CoreLab Partners, Inc conducted the reader study component of this investigation. They provided the reading facility, review workstations, software, and logistical support. CoreLab Partners radiologists also participated as readers. Therefore, we acknowledge CoreLab Partners for their support and specifically acknowledge CoreLab Partners radiologists Kevin Byrne, Steven Kaplan, Julie Barudin, Joyce Sherman, Kathy Slazak, George Edeburn, and J. Michael O'Neal for participating as readers in this study. Finally, we acknowledge financial support from the RSNA Quantitative Imaging Biomarker Alliance (QIBA) provided by National Institute of Biomedical Imaging and BIoengineering American Recovery and Reinvestment Act of 2009 funds. Certain commercial equipment, instruments, software, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the materials or equipment identified are necessarily the best available for the purpose. Similarly, the mention of commercial entities, or commercial products, their sources, or their use in connection with materials reported herein is not to be construed as either an actual or implied endorsement of such entities or products by the Department of Health and Human Services or the United States Food and Drug Administration.

Appendix A. A Sample Set of Results to Illustrate Reader Variability

An example to illustrate the sources of variation is based on the lesion displayed in Figure 1*D*. For this individual case, we have the results for both reading paradigms and all five readers for volumetric measurements shown in Table A1. This table demonstrates that even within *a single lesion* there are differences across both reading paradigm and readers. From this table, the averages of the absolute values of percent change are 101.8% for the independent reading and 15.9% for the locked sequential read. This table also shows the range of percent change values across readers: in the independent reading, the percent changes for different readers ranged from −16.4% to 240%, while for the locked sequential read, the percent changes for different readers ranged from −27.1% to 27.5%. As can be seen in Figure 1*D*, this lesion is reasonably complex and was not judged to be compliant with the QIBA CTV-1 profile, but in the locked sequential reads, the percent differences were within 30%. These results demonstrate the uncertainty of obtaining a consistent boundary for this lesion, resulting in variation from reader to reader that extends across both reading paradigms. Even under the locked sequential reading paradigm, where the readers did have access to their contours from the first time point scan, there is still uncertainty in the contours.

References

[1] Therasse P, Arbuck SG, and Eisenhauer EA, et al (2000). New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* **92**, 205–216.

[2] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, and Mooney M, et al (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* **45**(2), 228–247.

[3] Mozley PD, Bendtsen C, Zhao B, Schwartz LH, Thorn M, Rong Y, Zhang L, Perrone A, Korn R, and Buckler AJ (2012). Measurement of tumor volumes improves RECIST-based response assessments in advanced lung cancer. *Transl Oncol* **5**(1), 19–25.

[4] McNitt-Gray MF, Bidaut LM, Armato SG, Meyer CR, Gavrielides MA, Fenimore C, McLennan G, Petrick N, Zhao B, and Reeves AP, et al (2009).

Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error. *Transl Oncol* **2**(4), 216–222.

[5] Meyer CR, Armato SG, Fenimore CP, McLennan G, Bidaut LM, Barboriak DP, Gavrielides MA, Jackson EF, McNitt-Gray MF, and Kinahan PE, et al (2009). Quantitative imaging to assess tumor response to therapy: common themes of measurement, truth data, and error sources. *Transl Oncol* **2**(4), 198–210.

[6] Petrick N, Kim HJ, Clunie D, Borradaile K, Ford R, Zeng R, Gavrielides MA, McNitt-Gray MF, Lu ZQ, and Fenimore C, et al (2014). Comparison of 1D, 2D, and 3D nodule sizing methods by radiologists for spherical and complex nodules on thoracic CT phantom images. *Acad Radiol* **21**(1), 30–40.

[7] Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, Qin Y, Riely GJ, Kris MG, and Schwartz LH (2009). Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non–small cell lung cancer. *Radiology* **252**(1), 263–272.

[8] Oxnard GR, Zhao B, Sima CS, Ginsberg MS, James LP, Lefkowitz RA, Guo P, Kris MG, Schwartz LH, and Riely GJ (2011). Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. *J Clin Oncol* **29**(23), 3114–3119.

[9] Armato III S, Meyer C, McNitt-Gray M, McLennan G, Reeves A, Croft B, and Clarke L (2008). The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: a resource for the development of change-analysis software. *Clin Pharmacol Ther* **84**(4), 448–456 [Image data accessible at https://public.cancerimagingarchive.net/ncia/login.jsf and https://wiki.cancerimagingarchive.net/display/Public/RIDER+Collections; both accessed September 22, 2014].

[10] CT Volumetry Technical Committee (2012). CT Tumor Volume Change (CTV-1), Quantitative Imaging Biomarkers Alliance. Version 2.2. Reviewed Draft. QIBA; 2012 [Available from http://rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/QIBA_CT%20Vol_TumorVolumeChangeProfile_v2.2_PubliclyReviewedVersion_08AUG2012.pdf (accessed September 22, 2014)].

[11] Bankier A, Levine D, Halpern EF, and Kressel HY (2010). Consensus interpretation in imaging research: is there a better way? *Radiology* **257**(1), 14–17.

[12] Buckler AJ, Mozley PD, Schwartz L, Petrick N, McNitt-Gray M, Fenimore C, O'Donnell K, Hayes W, Kim HJ, and Clarke L, et al (2010). Volumetric CT in lung cancer: an example for the qualification of imaging as a biomarker. *Acad Radiol* **17**(1), 107–115.

[13] Jolly MP and Grady L (2008). 3D general lesion segmentation in CT. 5th IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro; 2008. p. 796–799.

[14] Grady L (2006). Random walks for image segmentation. *IEEE Trans Pattern Anal Mach Intell* **28**(11), 1768–1783.

[15] Bland JM and Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **327**(8476), 307–310.

[16] Laird N, Lange N, and Stram D (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *J Am Stat Assoc* **82**, 97–105.

[17] Kessler LG, Barnhart HX, and Buckler AJ, et al (2014). The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Stat Methods Med Res* [pii: 0962280214537333, Epub ahead of print].

[18] Obuchowski NA, Barnhar HX, and Buckler AJ, et al (2014). Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example. *Stat Methods Med Res* [pii: 0962280214537392, Epub ahead of print].

[19] Chow SC, Shao J, and Wang H (2003). Sample Size Calculations in Clinical Research. New York: Marcel Dekker; 2003 41–73.